Conference Applications of Mathematics 2012
in honor of the 60th birthday of Michal Křížek.
Institute of Mathematics AS CR, Prague 2012

THE CONVERGENCE OF EXPLICIT RUNGE–KUTTA METHODS COMBINED WITH RICHARDSON EXTRAPOLATION

István Faragó[1], Ágnes Havasi[2], Zahari Zlatev[3]

[1] Department of Applied Analysis and Computational Mathematics, Eötvös Loránd University, Budapest, 1117 Budapest, Pázmány P. s. 1/C, Hungary
faragois@cs.elte.hu

[2] Department of Meteorology, Eötvös Loránd University, Budapest
1117 Budapest, Pázmány P. s. 1/A, Hungary
hagi@nimbus.elte.hu

[3] National Environmental Research Institute, Aarhus University
Roskilde, Denmark
zz@dmu.dk

Abstract

Runge–Kutta methods are widely used in the solution of systems of ordinary differential equations. Richardson extrapolation is an efficient tool to enhance the accuracy of time integration schemes. In this paper we investigate the convergence of the combination of any explicit Runge–Kutta method with active Richardson extrapolation and show that the obtained numerical solution converges under rather natural conditions.

1. Introduction

This paper is concerned with the numerical solution of initial value problems of the form

$$y' = f(x, y), \quad y(a) = \eta, \tag{1.1}$$

where $y : \mathbb{R} \to \mathbb{R}^m$ is the unknown vector function, $f : \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^m$ and $\eta \in \mathbb{R}^m$ is a given initial vector. A solution is sought on the interval $[a, b]$ of x, where a and b are finite. It is assumed that f satisfies a Lipschitz condition, so that there existst a unique solution $y(x)$ of (1.1).

Explicit Runge–Kutta methods have the general form

$$y_{n+1} = y_n + h \sum_{i=1}^{m} b_i k_i \tag{1.2}$$

with

$$k_1 = f(x_n, y_n), \quad k_i = f\left(x_n + c_i h, y_n + h \sum_{j=1}^{i-1} a_{ij} k_j\right), \quad i = 2, \ldots, m, \tag{1.3}$$

where b_i, c_i and $a_{ij} \in \mathbb{R}$ are given constants. Here, as usual, y_n denotes an approximation to the solution $y(x_n)$ of (1.1) at x_n.

Richardson extrapolation is a powerful tool to increase the accuracy of some numerical method. It consists in applying the given numerical scheme with different discretization parameters (usually, h and $h/2$) and combining the obtained numerical solutions by properly chosen weights. Namely, if p denotes the order of the selected numerical method, w_n the numerical solution obtained by $h/2$ and z_n that obtained by h, then the combined solution

$$y_n = \frac{2^p w_n - z_n}{2^p - 1}$$

has order $p + 1$. This method was first extensively used by L. F. Richardson, who called it "the deferred approach to the limit" [7]. The Richardson extrapolation is especially widely used for time integration schemes, when, as a rule, the results obtained by two different time-step sizes are combined.

The Richardson extrapolation can be implemented in two different ways when one attempts to increase the accuracy of a time integration method. When the active Richardson extrapolation is used, the improved approximation for a given time layer is not used in the further computations, while it is used in the computation of the next approximation when the active Richardson extrapolation is utilized, see [9] in more detail. These two version of the Richardson extrapolation are also described in [2], where they are called global and local Richardson extrapolations.

It is not difficult to see that if the passive device is applied and the underlying method has some qualitative properties (e.g., it is stable / convergent), then the combined method also possesses property.

However, if the active device is used, then this is not valid anymore: any property of the underlying method does not imply the same property of the combined method. Therefore, the active Richardson extrapolation requires further investigation when a given numerical method is applied. (That is why in the sequel we will focus on the active version, and so Richardson extrapolation should always understood as active Richardson extrapolation.) So far the studies have been concerned with the order of the combined method, see e.g., [5], and its applications, such as air pollution modelling [4], the Maxwell equations [2] and diffusion-convection equations with large gradients [1].

During the applications, the investigation of A-stability of the combined method is of great importance, therefore this issue has been widely investigated in the previous works. In [4], the stability of the Richardson extrapolation combined with the backward Euler method and the trapezoidal rule was studied and applied efficiently in an atmospheric chemistry model. In [9] the stability of the Richardson extrapolation combined with the general θ-method was studied in detail. It is important to emphasize that these papers are concerned with the study of A-stability, which characterizes the behavior of the numerical method on Dahlquist's test problem [3], and on a fixed mesh.

In this paper we concentrate on the question of convergence, which, according to our knowledge, has not been investigated, and it was always hiddenly assumed in the works. Here the question is whether the numerical solution converges to the exact solution by reducing the step size. As we will see, this requires the property of the well-known zero-stability. (We remind the reader of the basic difference between A-stability and zero-stability: the first one gives the characterization of the numerical method on some fixed mesh, while the second one examines the method on the sequence of meshes with mesh sizes tending to zero.)

For the proof of the convergence we refer to a fundamental theorem of this subject, see [6], p. 36, which we cite here. Consider the numerical method written in the general form

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h\Phi_f(y_{n+k}, y_{n+k-1}, \ldots, y_n, x_n; h), \tag{1.4}$$

where the subscript f on the right-hand side indicates that the dependence of Φ on $y_{n+k}, y_{n+k-1}, \ldots, y_n, x_n$ is through the function $f(x, y)$. We impose the following two conditions on (1.4):

$$\left.\begin{aligned} \Phi_{f\equiv 0}(y_{n+k}, y_{n+k-1}, \ldots, y_n, x_n; h) &\equiv 0, \\ \|\Phi_f(y_{n+k}, y_{n+k-1}, \ldots, y_n, x_n; h) - \Phi_f(y^*_{n+k}, y^*_{n+k-1}, \ldots, y^*_n, x_n; h)\| & \\ \leq M \sum_{j=1}^{k} \|y_{n+j} - y^*_{n+j}\|, & \end{aligned}\right\} \tag{1.5}$$

where M is a constant. (These conditions are not very restrictive, e.g., the second one is automatically satisfied if the initial value problem to be solved satisfies a Lipschitz condition.)

Theorem 1.1. *The necessary and sufficient conditions for the method (1.4) to be convergent are that it be both consistent and zero-stable.*

The necessary and sufficient conditions for consistency can be expressed by the first characteristic polynomial $\rho(\zeta) = \sum_{j=0}^{k} \alpha_j \zeta^j$ of the method, namely, the method (1.4) is consistent iff

$$\rho(1) = 0 \tag{1.6}$$

and

$$\Phi_f(y(x_n), y(x_n), \ldots, y(x_n), x_n; 0)/\rho'(1) = f(x_n, y(x_n)), \tag{1.7}$$

see [6], p. 30.

For the condition of zero-stability we refer to the theorem on p. 35 of the same book:

Theorem 1.2. *The necessary and sufficient condition for the method (1.4) to be zero-stable is that it satisfies the root condition, i.e., the roots of ρ have modulus less than or equal to unity, and those of modulus unity are simple.*

Now our task is to write the combination of the explicit Runge–Kutta method and Richardson extrapolation in the form of (1.4), and show that it possesses (1.5), (1.6), (1.7) and the root condition of zero-stability.

2. The combined method as a one-step numerical method

The combination of the general explicit Runge–Kutta method with the Richardson extrapolation can be constructed in the following steps:

1) Make one step by time step h:

$$\begin{aligned} y_{n+1}^{(1)} &= y_n + h\sum_{i=1}^{m} b_i k_i, \\ k_1 &= f(x_n, y_n) \\ k_i &= f\left(x_n + c_i h, y_n + h\sum_{j=1}^{i-1} a_{ij} k_j\right) \end{aligned}$$

2) Make a step by time step $h/2$:

$$y_{n+\frac{1}{2}} = y_n + \frac{h}{2}\sum_{i=1}^{m} b_i \tilde{k}_i,$$

where

$$\begin{aligned} \tilde{k}_1 &= f(x_n, y_n) \\ \tilde{k}_i &= f\left(x_n + c_i\frac{h}{2}, y_n + \frac{h}{2}\sum_{j=1}^{i-1} a_{ij}\tilde{k}_j\right). \end{aligned}$$

From the obtained solution make a further step by $h/2$:

$$y_{n+1}^{(2)} = y_{n+\frac{1}{2}} + \frac{h}{2}\sum_{i=1}^{m} b_i \tilde{\tilde{k}}_i,$$

where

$$\begin{aligned} \tilde{\tilde{k}}_1 &= f\left(x_n + \frac{h}{2}, y_{n+\frac{1}{2}}\right) \\ \tilde{\tilde{k}}_i &= f\left(x_n + \frac{h}{2} + c_i\frac{h}{2}, y_{n+\frac{1}{2}} + \frac{h}{2}\sum_{j=1}^{i-1} a_{ij}\tilde{\tilde{k}}_j\right). \end{aligned}$$

By computing a weighed average of the results by using the weights d_1 for the solution obtained by h and d_2 for that obtained by $h/2$ ($d_1+d_2=1$), the combined numerical solution reads

$$y_{n+1} = d_1 y_{n+1}^{(1)} + d_2 y_{n+1}^{(2)} = d_1 \left[y_n + h \sum_{i=1}^{m} b_i k_i \right] + d_2 \left[y_n + \frac{h}{2} \sum_{i=1}^{m} b_i (\tilde{k}_i + \tilde{\tilde{k}}_i) \right].$$

From this, making use of the equality $d_1 y_n + d_2 y_n = y_n$, we obtain

$$y_{n+1} - y_n = d_1 h \sum_{i=1}^{m} b_i k_i + d_2 \frac{h}{2} \sum_{i=1}^{m} b_i (\tilde{k}_i + \tilde{\tilde{k}}_i).$$

So, the function $\Phi_f =: \Phi_f^{RE}$ corresponding to the combined method has the form

$$\Phi_f^{RE}(y_n, x_n; h) = d_1 \sum_{i=1}^{m} b_i k_i + \frac{d_2}{2} \sum_{i=1}^{m} b_i (\tilde{k}_i + \tilde{\tilde{k}}_i). \tag{2.8}$$

3. Checking the conditions for consistency and zero-stability

We have seen that the combined method has the form (1.4), where $k=1$ (one-step method), $\alpha_0 = -1,\ \alpha_1 = 1$ and Φ_f is as under (2.8). As one can easily check, the first characteristic polynomial of the method is $\rho(\zeta) = -1 + \zeta$.

First we show that (1.5) holds under the usual conditions for the IVP. The first condition follows from the fact that if $f \equiv 0$, then all the functions $k_i, \tilde{k}_i$ and $\tilde{\tilde{k}}_i$ are identically zero. It remains to check the Lipschitz condition.

It is sufficient to show that $k_i, \tilde{k}_i$ and $\tilde{\tilde{k}}_i$ satisfy a Lipschitz condition, provided that so does f, i.e.,

$$|f(x_n, y_n) - f(x_n, y_n^*)| \le L |y_n - y_n^*|.$$

From the Lipschitz property of f it follows that k_1 and $\tilde{k}_1$ also satisfy this property with Lipschitz constant L.
Denote $a_{\max} = \max_{i,j} |a_{ij}|$. Then

$$\begin{aligned} |k_i - k_i^*| &= \left| f\left(x_n + c_i h, y_n + h \sum_{j=1}^{i-1} a_{ij} k_j \right) - f\left(x_n + c_i h, y_n^* + h \sum_{j=1}^{i-1} a_{ij} k_j^* \right) \right| \\ &\le L \left| y_n + h \sum_{j=1}^{i-1} a_{ij} k_j - y_n^* - h \sum_{j=1}^{i-1} a_{ij} k_j^* \right| \le L |y_n - y_n^*| + Lh \sum_{j=1}^{i-1} a_{ij} |k_j - k_j^*| \\ &\le L \left[|y_n - y_n^*| + h a_{\max} \sum_{j=1}^{i-1} |k_j - k_j^*| \right]. \end{aligned}$$

Assume that $k_1, k_2, k_3, \ldots, k_{i-1}$ all satisfy a Lipschitz condition, i.e.,

$$|k_l - k_l^*| \le L_l |y_n - y_n^*|, \quad l = 1, 2, \ldots, i-1.$$

Then

$$
\begin{aligned}
|k_i - k_i^*| &\le L \left[|y_n - y_n^*| + h a_{\max} \sum_{j=1}^{i-1} |k_j - k_j^*| \right] \\
&\le L \left[|y_n - y_n^*| + h a_{\max} \left(\sum_{j=1}^{i-1} L_j \right) |y_n - y_n^*| \right] \\
&\le L \left[1 + (b-a) a_{\max} \sum_{j=1}^{i-1} L_j \right] |y_n - y_n^*|,
\end{aligned}
$$

where we used the length of the interval $[a, b]$ as an upper bound for h. So, k_i satisfies a Lipschitz condition with Lipschitz constant $L_i = L[1 + (b - a)a_{\max} \sum_{j=1}^{i-1} L_j]$. The constant L_i can be expressed by L_{i-1} as

$$
\begin{aligned}
L_i &= L \left[1 + (b-a) a_{\max} \left(\sum_{j=1}^{i-2} L_j + L_{i-1} \right) \right] \\
&= L_{i-1} + L(b-a) a_{\max} L_{i-1} = (1 + L(b-a) a_{\max}) L_{i-1}.
\end{aligned}
$$

Consequently, $L_i = (1 + L(b-a)a_{\max})^{i-1} L_1 = (1 + L(b-a)a_{\max})^{i-1} L$.

Since $\tilde{k}_1 = k_1$, therefore the same holds for $\tilde{k}_i$.

Finally, the Lipschitz property of $\tilde{\tilde{k}}_i$ follows from that of $\tilde{k}_i$, since

$$
\begin{aligned}
|\tilde{\tilde{k}}_1 - \tilde{\tilde{k}}_1^*| &= \left| f\left(x_n + \frac{h}{2}, y_n + \frac{h}{2} \sum_{i=1}^{m} b_i \tilde{k}_i \right) - f\left(x_n + \frac{h}{2}, y_n^* + \frac{b-a}{2} \sum_{i=1}^{m} b_i \tilde{k}_i^* \right) \right| \\
&\le L|y_n - y_n^*| + L \frac{h}{2} \sum_{i=1}^{m} b_i |\tilde{k}_i - \tilde{k}_i^*|
\end{aligned}
$$

$$
\begin{aligned}
|\tilde{\tilde{k}}_i - \tilde{\tilde{k}}_i^*| =& |f\left(x_n + \frac{h}{2} + \theta_i \frac{h}{2}, y_n + \frac{h}{2} \left(\sum_{i=1}^{m} b_i \tilde{k}_i + \sum_{j=1}^{i-1} a_{ij} \tilde{\tilde{k}}_j \right) \right) - \\
& - f\left(x_n + \frac{h}{2} + c_i \frac{h}{2}, y_n^* + \frac{h}{2} \left(\sum_{i=1}^{m} b_i \tilde{k}_i^* + \sum_{j=1}^{i-1} a_{ij} \tilde{\tilde{k}}_j^* \right) \right) \\
\le & L|y_n - y_n^*| + L \frac{b-a}{2} \left(\sum_{i=1}^{m} b_i |\tilde{k}_i - \tilde{k}_i^*| + \sum_{j=1}^{i-1} a_{ij} |\tilde{\tilde{k}}_j - \tilde{\tilde{k}}_j^*| \right).
\end{aligned}
$$

Now we check the conditions of consistency, i.e., (1.6) and (1.7). The first one is easy to see by substituting $\zeta = 1$ into the first characteristic polynomial $\rho(\zeta) = -1 + \zeta$. Since $\rho'(1) = 1$, therefore the second condition reduces to the equality

$$
\Phi_f^{RE}(y(x_n), x_n; 0) = f(x_n, y(x_n)). \tag{3.9}
$$

For the combined method we have

$$\Phi_f^{RE}(y_n, x_n; 0) = d_1 \sum_{i=1}^{m} b_i f(x_n, y_n) + \frac{d_2}{2} \left(\sum_{i=1}^{m} b_i \right) 2f(x_n, y_n)$$
$$= (d_1 + d_2) \left(\sum_{i=1}^{m} b_i \right) f(x_n, y_n) = f(x_n, y_n),$$

which holds if and only if $\sum_{i=1}^{m} b_i = 1$, which is always assumed, because it is required for the consistency of the underlying Runge–Kutta method.

It remains to show that zero-stability holds for the combined method. According to Theorem 1.2., the root condition is to be checked. Now the only root of the first characteristic polynomial $\rho(\zeta) = -1 + \zeta$ is equal to unity, therefore the combined method trivially satisfies the root condition, and so the method is zero-stable.

Hence, we have proven the main result of the paper.

Theorem 3.1. *Assume that some explicit Runge–Kutta method combined with the active Richardson extrapolation is applied to problem (1.1), satisfying a Lipschitz condition. Then the combined method is convergent.*

4. Conclusion

In this paper we have shown that the combination of any explicit Runge–Kutta method with the (active) Richardson extrapolation results in a convergent numerical method under some rather natural conditions. In the proof we have used the concepts of consistency and zero-stability.

In the future we plan to investigate the combination of a wider group of methods, the so-called diagonally implicit Runge–Kutta methods [8] in combination with the Richardson extrapolation.

Acknowledgements

The European Union and the European Social Fund have provided financial support to the project under the grant agreement no. TÁMOP 4.2.1./B-09/1/KMR-2010-0003.

References

[1] Andreev, V. F. and Popov, A. M.: Using Richardson's method to construct high-order accurate adaptive grids. Comput. Math. Model. **10** (1999), 227–238.

[2] Botchev, M. A. and Verwer, J. G.: Numerical integration of damped Maxwell equations. SIAM J. Sci. Comput. **31** (2009), 1322–1346.

[3] Dahlquist, G.: A special stability problem for linear multistep methods. BIT **3** (1963), 27–43.

[4] Faragó, I., Havasi, Á., and Zlatev, Z.: Efficient implementation of stable Richardson Extrapolation algorithms. Comput. Math. Appl. **60** (2010), 2309–2325.

[5] Geiser, J.: Higher-order difference and higher-order splitting methods for 2D parabolic problems with mixed derivatives. International Mathematical Forum **2** (2007), 3339–3350.

[6] Lambert, J. D.: *Numerical methods for ordinary differential equations*. Wiley, New York, 1991.

[7] Richardson, L. F.: The deferred approach to the limit, I-single lattice. Philosophical Transactions of the Royal Society of London **226** (1927), 299–349.

[8] Zlatev, Z.: Modified diagonally implicit Runge-Kutta methods. SIAM Journal on Scientific and Statistical Computing **2** (1981), 321–334.

[9] Zlatev, Z., Faragó , I., and Havasi, Á.: Stability of the Richardson Extrapolation applied together with the θ-method. J. Comput. Appl. Math. **235** (2010), 507–517.